Scientific and Technical Report

Sponsored by
Advanced Research Projects Agency/ITO
and United States Patent and Trademark Office

Browsing, Discovery and Search in Large Distributed Databases
of Complex and Scanned Documents

ARPA Order No. D570

Issued by EXC/AXS under Contract #F19628-95-C-0235

Date Submitted: April 9, 1998

Period of Report: January 1, 1998 to March 31, 1998

Submitted by: Professor W. Bruce Croft, Principal Investigator
Computer Science Department
University of Massachusetts, Amherst

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	04/09/98	Scientific/Tech

4. TITLE AND SUBTITLE
Browsing, Discovery, and Search in Large Distributed Databases of Complex and Scanned Documents

5. FUNDING NUMBERS
F19628-95-C-0235
ARPA Order No. D570

6. AUTHOR(S)
W. Bruce Croft

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Massachusetts, Amherst
Box 36010, OGCA, Munson Hall
Amherst, MA 01003-6010

8. PERFORMING ORGANIZATION REPORT NUMBER
TR5281810498

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Mr. Harry Koch
ESC/AXS
Bldg 1704. Room 114
5 Eglin St.
Hanscom AFB, MA 01731-2116

Ms. Monique Dillon
Office of Naval Research
Boston Regional Office
495 Summer St., Room 103
Boston, MA 02210-2109

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution Statement A: Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This project aims to integrate powerful, new techniques for interactive browsing, discovery, and retrieval in very large, distributed databases of complex and scanned documents. Emphasis is placed on going beyond full-text retrieval techniques developed in the DARPA TIPSTER program to support different types of access and non-textual content. These techniques should be particularly relevant to the patent domain where it is important to find relationships between documents and where the patent or trademark may be based on a visual design. The specific tasks identified involve studying representation techniques for long documents with complex structure, browsing and discovery techniques for large text databases, image retrieval and scanned document retrieval techniques, and architectures for large, distributed databases.

14. SUBJECT TERMS

Browsing	Query Processing	Indexing
Image Retrieval	Scanned Document Retrieval	Bayesian Network
Text Retrieval	Probabilistic Retrieval Model	Large Distributed Databases

15. NUMBER OF PAGES
11

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

Table of Contents

Browsing, Discovery and Search in Large Distributed Databases of Complex and Scanned Documents

Technical and Scientific Report

Task 1: Representation Techniques for Complex Documents

Task Objectives

In this task, the goal is to extend the word-based representations that are common in retrieval systems in order to support summarization, browsing, and more effective retrieval. Specifically, we will be studying phrase-based representations and relationships between phrases in individual and groups of documents as the basis for our approach. Document structure will be used as part of the information that is used to "tag" the phrasal representation.

Technical Problems

The technical problems have to do with defining a "phrase", developing techniques for rapidly extracting them from text, comparing phrase contexts to identify significant relationships, producing summaries from these representations, extending the underlying retrieval model to be able to make effective use of phrasal representations, and using complex document structure in indexing and retrieval.

General Methodology

The general methodology for this task is to demonstrate effectiveness through user-based and collection-based experiments. As well as the PTO text databases, we will make extensive use of the TIPSTER document collection, which consists of a large number of text documents from a variety of sources, queries, and user relevance judgments for each query.

Technical Results

We continued to develop the use of phrases from the PTO text. We are currently focusing on developing a new technique for phrase identification in the indexing process. The list of candidate phrases identified by the lexical acquisition program is very large, and using this list during indexing requires substantial memory and computational resources. As an alternative approach, we are attempting to train a phrase recognizer using the list, similar to the Markov Model techniques that have been used for part of speech tagging. If we are

successful, phrase recognition will be faster, require less memory, and will be able to recognize new phrases that are not in the vocabulary. Preliminary results are encouraging.

Progress was made on incorporating query expansion into the patent search.

Important Findings and Conclusions

Our experiments showed that phrase indexing and query formulation techniques can substantially improve the results of patent searches. Further experiments are required to evaluate the new phrase recognition approach.

Significant Hardware Development

None

Special Comments

None.

Implication for Further Research

We plan further testing of the phrase recognition approach. We also plan to spend more time with patent searchers to understand how best to combine Boolean and free text queries at the interface level.

Task 2: Browsing and Classification Techniques for Document Collections

Task Objectives

The goals of this task are to develop techniques for summarizing and classifying collections of documents. These techniques will be designed to support interactive browsing and text classification in environments like the PTO.

Technical Problems

The technical problems involve producing an effective summary of a group of documents, such as a retrieved set or an entire database. Both document and phrase clusters could be used as part of this process. The classification task emphasizes the ability to accurately assign predefined categories (as in the PTO classification) to new documents (patents). An additional problem is to determine when existing classifications do not match well to

new documents, such as when a PTO category covers too many patents and needs to be refined.

General Methodology

Evaluation of these techniques will be done using both the TREC corpus and PTO data. For the classification task in particular, we are designing evaluation criteria with substantial input from PTO staff.

Technical Results

A number of classification experiments were performed and additional datasets are being created to evaluate the classification approaches being used.

The "document placement test" from the PTO test scenario was a simple task which produced excellent results.

Based on experiments with k-nearest-neighbor classifiers for some randomly selected documents, we decided to train classifiers for a set of closely related subclasses that had sufficient numbers of examples in them. We are currently setting up a dataset which contains all the documents from 1985-1997 in the roughly 100 subclasses under the "speech signal processing" node in the PTO hierarchy. We plan to train classifiers to place documents in the 30 or so subclasses fall under the "speech recognition" node, using the others as closely related negative examples. The years 1985-1995 will be used as training data and 1996-1997 as test data.

A paper was accepted for the AAAI Workshop on Learning for Text Categorization, entitled "Some issues in the automatic classification of US Patents."

Experiments on the visualization of retrieval results continue. Two papers describing this work have recently been accepted.

R. Swan and J. Allan, "Aspect Windows, 3-D Visualizations, and Indirect Comparisons of Information Retrieval Systems", to be presented at SIGIR 98.

A. Leouski and J. Allan, "Visual Interactions with Multidimensional Ranked List", to be presented at SIGIR 98.

Important Findings and Conclusions

Classification results are too preliminary for conclusions, but the issues are discussed in the paper attached to this report. The papers on visualization contain discussions of a number of results.

Significant Hardware Development

None

Special Comments

None.

Implication for Further Research

We continue to focus on evaluating the classification accuracy, incorporating additional classification techniques into the classification system, and evaluating visualization techniques.

Task 3: Image Indexing and Retrieval

Task Objectives

The goal of this task is to develop similarity-based techniques for retrieving images such as trademarks, logos, and designs.

Technical Problems

The central issue is how images can be indexed to support efficient, content-based retrieval. The primary type of query in these environments is "find me things that look like this". We are developing "appearance-based" retrieval of images as well as more straightforward features such as color and texture. Filter based and frequency domain based techniques offer some potential in this area, but significant work needs to be done on making this approach efficient enough to deal with hundreds of thousands of images.

General Methodology

The evaluation of these techniques will be done in a similar way to text by developing test collections of images. Specifically, we are working to obtain large collections of trademark and design images, both from the PTO and from general sources such as the web.

Technical Results

Over the last year, we have developed whole image matching techniques based on histograms of curvature and phase at multiple scales (see [2,3]. A technique based on comparing jpeg coefficients of images was also developed. Text and image searching were combined for trademark searching, i.e. we used INQUERY for searching through the text database and searched images using one of the whole image techniques above.

The technique was applied to 60,000 trademark images and this work was demonstrated at the PTO/DARPA meeting in December. There were some bugs in the curvature/phase technique which have since been ironed out. We have also studied moment based techniques used in the literature and showed that they do not work as well for trademark searching as the techniques we have developed.

We are now working on adding indexing to the curvature/phase technique. This will allow us to add new images for searching at query time.

We have recently finished building a database for the complete trademarks (654,835 images). We are currently waiting on some additional data from the PTO (design code data) to finish building the database to demonstrate next quarter.

We have some new work on detecting flowers in plant patents. First, the images are separated from the text on the patent page. Then a flower is detected by segmenting on the basis of color. The system exploits domain knowledge to do this, e.g. flowers are usually not green, black or brown in color.

The segmented flowers are then retrieved using color or the names of colors. We have done some work with a small database of about 100 flower patents as well as a 150 other images taken from other sources. At present, the database is too small to effectively test the method. We are expecting more flower data from the PTO for this project.

We have also finished porting the system for detecting text from images to a standalone system. We are now working on finishing the debugging and further speeding up the system (Processing a 2000 by 3000 pixel image now takes about 10 min on a Pentium Pro 200 system). Studies comparing our document clean-up and binarization technique with others were also done.

The following papers were presented recently:

1) Wu, V. and Manmatha, R., " "Document Image Clean-up and Binarization"
in the Proceedings of SPIE conf. on Document Recognition V, San Jose, California, January 24 - 30, 1998.

2) Manmatha. R., Ravela, S. and Chitti Y., "On Computing Local and Global Similarity in Images" in the Proceedings of SPIE conf. on Human and Electronic Imaging III, San Jose, California, January 24 - 30, 1998.

The following papers have been submitted:

3) Ravela, S. and Manmatha, R., "On computing global similarity in images" submitted to the IEEE Workshop on Applications of Computer Vision (WACV"98), October 1998.

4) Das, M., Manmatha, R. and Riseman, E. M., "Indexing Flowers by Color Names using Domain Knowledge-driven Segmentation"

The following paper is under review for a journal:

5) Wu, V., Manmatha, R. and Riseman, E. M., "Finding Text in Images" under revision for the IEEE Transactions of Pattern Analysis and Machine Intelligence"

Important Findings and Conclusions

These are discussed in the papers attached to the report.

Significant Hardware Development

None

Special Comments

The progress of this part of the project depends on data from the PTO. Specifically, we need the design code data and more flower patents.

Implication for Further Research

We will focus on developing the large scale trademark retrieval demosntration.

Task 4: Distributed Retrieval Architecture

Task Objectives

The goals of this task are to scale up our current methods of automatically selecting collections and merging results, and to investigate architectures that can support efficient retrieval, browsing and relevance feedback in distributed environments with terabytes of information.

Technical Problems

The current INQUERY text retrieval system uses a client server architecture to support simultaneous retrieval from multiple collections distributed across one or more processors. A number of efficiency bottlenecks develop, however, when the size of the databases is very large. Deciding which subcollections to search can address part of the problem, but there are other problems associated with the fundamental efficiency of the processes involved and the use of distributed resources. Image indexing and retrieval tends to make all of these problems worse since the databases and indexes are considerably larger.

General Methodology

The architectures and algorithms produced in this task will be evaluated using a combination of standard performance (efficiency) measures and effectiveness measures. The efficiency tests will be done using TREC data and large PTO databases, including images, and the collection selection algorithms will be evaluated using the text subcollections of the patents.

Technical Results

During this quarter a new version of the patent search demo was created that uses Inquery's multi-database (MDB) capabilities. The new version automatically broadcasts each search request to every database, and then merges the returned results into a single result list. This configuration enables a large database to be distributed among multiple computers, improving response time for each search.

The demo runs over the AAInet, using two years of ASCII patent data. About 5 gigabytes of 1995 patents (ASCII form) are stored at UMass. Another 5 gigabytes of 1996 patents (ASCII form) are stored at the San Diego Supercomputer Center (SDSC). Each search request is broadcast, transparently, to each database. The results are merged, and shown as a single list of results. Only the database column in the GUI indicates which database supplied each document.

The initial implementation suffered from speed problems, because it needed to download each document in order to construct the list of retrieved documents. This problem was eliminated largely by more careful use of existing Inquery MDB API calls. The demo now only downloads a copy of the document when the user asks to see it.

The demo can be accessed at

http://darwin/cgi-bin/pto/pto_query_mdb/dummy

We have also continued experiments on new techniques for database selection in distributed search. This is reported in the following paper:

J. Xu and J. Callan, "Effective Retrieval with Distributed Collections", to appear in SIGIR 98.

Important Findings and Conclusions

The InQuery system appears to have the potential for scaling to terabyte-sized databases, but further experiments are required.

Significant Hardware Development

None.

Special Comments

SDSC has indicated that it will discontinue support of the AAInet at the end of May. The multi-database demo can be used after that time via an ordinary Internet connection, although each search will take more time, due to network congestion and delays.

Implications for Further Research

The current multi-database demo does not have all of the capabilities of the single database demo. For example, relevance feedback has not been implemented, although there is no serious obstacle to doing so. The demo could also be modified to be more selective about which databases it searches, although this only makes sense if the data is divided into a larger number of databases.